

Beschreibung

Die Erfindung betrifft ein Verfahren zur Spracherkennung durch einen Rechner.

5 Spracherkennung beginnt bei der Erfassung eines Sprachsignals in Form einer Abtastung und Digitalisierung. Es folgt eine Merkmalsextraktion (Ververarbeitung) und eine Klassifikation (auch: Distanzberechnung) des vverarbeiteten Sprachsignals. Im letzten Schritt der Spracherkennung (Suche) werden Wortmodelle zugeordnet, es ergibt sich eine erkannte Wortsequenz (siehe [1]).

Die Distanzberechnung stellt den rechenintensivsten Schritt bei der Spracherkennung dar. Je nach Anwendungsfall kann dabei der Anteil der benötigten Rechenleistung für die Distanzberechnung bis zu über 80% des gesamten Berechnungsaufwands des Spracherkenners betragen.

10 In [2] ist ein Verfahren zur Distanzberechnung genannt, das gegenüber der Erfindung den Nachteil aufweist, daß statistische Cluster entsprechend statistischer Verteilungen angeordnet sind, somit die Güte der Spracherkennung direkt von der Anzahl der Cluster abhängt und der Berechnungsaufwand hoch ist.

Aus [3] sind Hidden-Markov-Modelle (HMMs) zur Laut- und Wortmodellierung bekannt. Weiterhin sind in [4] 15 kontinuierliche mehrdimensionale, meist multivariate, Verteilungen (z. B. Laplace- oder Gauß-Verteilungen) für die Modellierung der Merkmale implementiert.

Die Aufgabe der Erfindung besteht darin, ein Verfahren zur Distanzberechnung anzugeben, das eine schnelle Klassifikation eines Merkmalsvektors (mit vorgebbarer Genauigkeit) ermöglicht.

Diese Aufgabe wird gemäß den Merkmalen des Patentanspruchs 1 gelöst.

20 Zu diskreten Abtastzeitpunkten wird ein Sprachsignal mit einer vorgebbaren Taktrate abgetastet und digitalisiert. Zu jedem Abtastzeitpunkt wird das Sprachsignal in einen Merkmalsvektor mit vorgebbarer Dimension abgebildet. Zur Durchführung der Distanzberechnung werden Abstände des jeweiligen Merkmalsvektors zu vorgebbaren Mustervektoren bestimmt.

Die Mustervektoren stellen unterscheidbare Lauteinheiten der Sprache dar. Jede Lauteinheit verfügt über 25 einen Kontextgrad, der angibt, wie weit der Kontext der Lauteinheit reicht.

Zunächst wird zu ersten Mustervektoren, die Lauteinheiten mit einem vorgebbaren ersten Kontextgrad darstellen, jeweils ein Abstand zu dem Merkmalsvektor berechnet. Aus all den so berechneten ersten Abständen wird ein erstes Minimum der ersten Abstände ausgewählt.

30 Um eine genauere Klassifikation des Merkmalsvektors vornehmen zu können, werden in einem nächsten Schritt zweite Mustervektoren, die Lauteinheiten mit einem zweiten Kontextgrad darstellen, wobei der zweite Kontextgrad größer als der erste Kontextgrad ist, innerhalb eines vorgebbaren Bereichs um die durch das erste Minimum klassifizierte Lauteinheit zur Berechnung von zweiten Abständen herangezogen. Ein zweites Minimum der zweiten Abstände klassifiziert eine Lauteinheit mit einem gegenüber der zu dem ersten Minimum gehörenden Lauteinheit höherem Kontextgrad.

35 Reicht die Güte der durchgeführten Klassifikation des Merkmalsvektors aus, so kann das Verfahren beendet werden, indem Sprache aus klassifizierten Lauteinheiten zusammengesetzt und somit erkannt wird.

Soll eine genauere Klassifikation erfolgen, so wird nochmals der Kontextgrad der zu untersuchenden Lauteinheiten erhöht, und somit in einem vorgebbaren Bereich um die zuletzt das zuletzt klassifizierte Minimum 40 Abstände zu dem Merkmalsvektor berechnet. Diese Iteration kann beliebig oft wiederholt werden, um ein immer besseres Ergebnis zu erhalten. Allerdings genügt eine Klassifikation mit hohem Kontextgrad nicht mehr der Forderung nach einer Echtzeitbedingung, da die Erhöhung des Kontextgrads den Aufwand für die Berechnung der Klassifikation typischerweise exponentiell ansteigen läßt.

Eine Weiterbildung des erfindungsgemäßen Verfahrens besteht darin, bei Verwendung von Lauteinheiten mit höherem Kontextgrad den nächsthöheren Kontextgrad auszuwählen.

45 Die Erfindung kann dahingehend weitergebildet werden, daß zu Beginn des Verfahrens Lauteinheiten mit dem Kontextgrad "1" verwendet werden.

Weiterhin können als solche Lauteinheiten zu Beginn des erfindungsgemäßen Verfahrens Phoneme oder Monophone einer Sprache verwendet werden.

50 Eine andere Weiterbildung der Erfindung besteht darin, Diphone oder Triphone oder aus mehreren Phonemen zusammengesetzte Laute als Lauteinheiten zu verwenden.

In den Figuren sind Ausführungsbeispiele der Erfindung dargestellt, die im weiteren näher erläutert werden.

Es zeigen

Fig. 1 ein Blockdiagramm, das Schritte des erfindungsgemäßen Verfahrens enthält,

Fig. 2 eine Skizze, die die Erfindung im zweidimensionalen Merkmalsraum veranschaulicht.

55 In Fig. 1 werden Schritte des erfindungsgemäßen Verfahrens für die Berechnung der Distanz einer Untereinheit dargestellt.

Vektoren werden nachfolgend als kleine fettgedruckte Zeichen oder mit einem Pfeil, Matrizen als große fettgedruckte Zeichen dargestellt.

60 Gesprochene Sprache wird abgetastet und in digitalisierter Form einem Rechner zugänglich gemacht. In Schritt 1a, der Vorverarbeitung, wird demzufolge in regelmäßigen Abstand (z. B. alle 10ms) ein Merkmalsvektor x , der das aktuelle Sprachsignal repräsentiert, der Dimension D_x (typische Werte für die Dimension: 20—64) ermittelt.

65 Wörter werden aus Einheiten modelliert, die eine Abfolge von Zuständen in einem Hidden-Markov-Modell (siehe ausführliche Beschreibung in [3]) darstellen. Solche Einheiten sind zumeist Phoneme, es können aber auch Halbsilben, Silben oder andere Einheiten verwendet werden. Eine Gesamtanzahl von Untereinheiten N_u , die die kleinsten sich unterscheidenden Einheiten von Sprache darstellen, schwankt je nach System und gewünschter Applikation (z. B. Einzelworterkennung, kontinuierliche Sprache, Sprecherunabhängigkeit) zwischen weniger als 100 und mehreren 1000 Untereinheiten N_u .

Bei der Distanzberechnung (siehe Fig. 1, Schritte 1b bis 1f) geht es darum, für alle Untereinheiten N_u akustische Distanzen gemäß einer verwendeten modellierenden Verteilung zu berechnen. Für jede Untereinheit u werden $N_v(u)$ Verteilungen zur Modellierung benutzt, wobei sich $N_v(u)$ für jede Untereinheit u unterscheiden kann. Typische Werte für $N_v(u)$ liegen zwischen 1 und 100. Bei Verwendung von Gauß-Verteilungen werden diese dargestellt als

$$g_i = G(\bar{\mu}_j, C_j) \quad (1),$$

wobei $\bar{\mu}_j$ der D_x -dimensionale Mittelpunktvektor und C die D_x -dimensionale Kovarianzmatrix der mit j indizierten Gauß-Verteilung g_j bezeichnen.

Die Gesamtverteilung für eine Untereinheit u (Summe der Gauß-Verteilungen) wird dargestellt durch g_u . Es handelt sich dabei um die mit c_j gewichtete Summe der einzelnen Verteilungen:

$$g_u = \sum_{j=1}^{N_v(u)} c_j \times g_j \quad (2).$$

Es ist üblich, von diagonalen Kovarianzmatrizen der Gaußverteilungen auszugehen. $\sigma_{j,i}$ bezeichnet dann die i -te Komponente der Varianz der Verteilung j . Die Wahrscheinlichkeit dafür, daß eine bestimmte Untereinheit u aufgetreten ist, wenn der beobachtete Merkmalsvektor x vorliegt, berechnet sich zu:

$$p_u(x) = k \times \sum_{j=1}^{N_v(u)} \frac{c_j}{\prod_{i=1}^{D_x} \sigma_{j,i}} \times \exp \left(-\frac{1}{2} \times \sum_{i=1}^{D_x} \frac{(x_i - \mu_{j,i})^2}{\sigma_{j,i}^2} \right) \quad (3),$$

wobei k eine vorgebbare implementierungsabhängige Konstante ist.

Eine verwendete Näherung geht davon aus, daß die Wahrscheinlichkeit einer Untereinheit u von der Wahrscheinlichkeit der "besten" (= wahrscheinlichsten) Verteilung dominiert wird. Dies bedeutet, daß in Gleichung (3) die äußere Summe durch einen max-Operator ersetzt werden kann. Es ist aber immer noch die Berechnung der Wahrscheinlichkeiten aller Verteilungen notwendig, denn erst dann kann der max-Operator ausgewertet werden:

$$p_u(x) = \max_j \left\{ k \times \frac{c_j}{\prod_{i=1}^{D_x} \sigma_{j,i}} \times \exp \left(-\frac{1}{2} \times \sum_{i=1}^{D_x} \frac{(x_i - \mu_{j,i})^2}{\sigma_{j,i}^2} \right) \right\} \quad (4).$$

$$\forall (j \in \{1, \dots, N_v(u)\})$$

Um die Berechnung des komplexen Terms in Gleichung (4) zu umgehen, wird Gleichung (4) logarithmiert, man spricht dann von Distanzen anstelle von Wahrscheinlichkeiten und bezeichnet diese Distanzen mit $d_u(x)$. Nach Umformen ergibt sich folgende Gleichung zur Berechnung der Distanz der akustischen Untereinheit u bei Vorliegen des betrachteten Merkmalsvektors x :

$$d_u(\mathbf{x}) = \min_j \left\{ K_{1,j} + K_{2,j} \times \sum_{i=1}^{D_x} (x_i - \mu_{j,i})^2 \right\}$$

(5),

$$\forall (j \in \{1, \dots, N_v(u)\})$$

wobei $K_{1,j}$ und $K_{2,j}$ von der Realisierung abhängige vorgebbare Konstanten darstellen, die einmal im Voraus berechnet werden können.

Ähnliche Berechnungsschemata ergeben sich auch bei Verwendung anderer Verteilungen, z. B. der Laplace-Verteilung, und schränken das erfindungsgemäße Verfahren nicht auf die ausgeführte Verteilungsvariante ein.

Gleichung (5) wird für alle Untereinheiten u ausgewertet. Die Gesamtzahl aller Verteilungen im System für alle Untereinheiten u sei N_v und berechnet sich zu

$$N_v = \sum_{k=1}^{N_u} N_v(k) \quad (6)$$

Das bedeutet, daß alle 10ms (alle 10ms wird ein Merkmalsvektor erstellt) N_v -mal ein D_x -dimensionaler Abstand berechnet wird, wobei N_v bei Systemen typischerweise im Bereich 1000–500 000 liegt. Diese Berechnung kann heutzutage von einem handelsüblichen Rechner (PC, Workstation, Signalprozessor) nicht unter Einhaltung der Echtzeitbedingung durchgeführt werden.

Das erfindungsgemäße Verfahren zeichnet sich dadurch aus, daß phonetisches Wissen zur Bestimmung einer Untermenge von Repräsentanten herangezogen wird. Es findet ein direktes Clustering von Untereinheiten statt.

Die Erfindung verwendet als ein primäres Kriterium zur Ermittlung einer geeigneten Untermenge zu berechnender Distanzen das Wissen um den phonetischen Kontext der modellierten (Sprach-)Segmente bzw. Untereinheiten.

Zur Erläuterung des erfindungsgemäßen Verfahrens wird von der Verwendung von Phonemen als Basis-Lauteinheiten für die Spracherkennung ausgegangen. Das Verfahren ist jedoch auf für andere Lauteinheiten als Phoneme realisierbar.

Ein typisches Spracherkennungssystem enthält ca. 40–60 Basis-Lauteinheiten. Bei solchen Basis-Lauteinheiten handelt es sich um kontextunabhängige Lauteinheiten. Um Sprache genauer klassifizieren zu können, werden Basis-Lauteinheiten in Abhängigkeit der vorangehenden und nachfolgenden Basis-Lauteinheiten gebildet.

Heutzutage werden Diphone, d. h. zwei Phoneme bestimmen die Untereinheit (ein Nachbarphonem wird zur Kontextbildung herangezogen), Triphone, d. h. drei Phoneme bestimmen die Untereinheit und Quinphone, d. h. fünf Phoneme bestimmen die Untereinheit, verwendet.

Kontext-abhängige Lauteinheiten werden wie folgt notiert:

$$L[i-1, \dots, i-1, i+1, \dots, i+r] \quad (7)$$

wobei i einen Zentrallaut markiert, bei dem 1 Laute im vorangehenden (= linken) Kontext und r Laute im nachfolgenden (= rechten) Kontext zur Bestimmung der Untereinheit berücksichtigt werden. Es liegt demnach ein $(1+r+1)$ -Laut vor.

Durch diese hier beschriebene Modellierung kann sich die Anzahl der Lauteinheiten mit höherem Kontextgrad, in Abhängigkeit der jeweiligen Anwendung auf mehrere 1000 erhöhen, da die Anzahl der verschiedenen theoretisch möglichen Lauteinheiten mit der Potenz der berücksichtigten Kontextlänge wächst.

Neben den kontext-abhängigen Lauteinheiten werden zusätzlich kontext-unabhängige Lauteinheiten (= Basis-Lauteinheiten) implementiert. Diese Basis-Lauteinheiten dienen bei "fastlook-ahead"-Methoden zur schnellen Vorauswahl von Hypothesen in der Suche nach einer Klassifikation eines Merkmalsvektors oder werden zur Modellierung von Wörtern herangezogen, wenn während einer Trainingsphase des Spracherkennungssystems zu wenige Ergebnisse vorlagen, um modellierende Verteilungen nach Gleichung (2) für die detaillierten Untereinheiten (im weiteren Kontext) zu bestimmen.

Für alle Lauteinheiten (Untereinheiten u) wird ein Kontextgrad G_k bestimmt, der angibt, wie weit der Kontext der jeweiligen Lauteinheit reicht. Für Basis-Lauteinheiten (kontextunabhängige Untereinheiten) gilt $G_k = 1$, für Diphone gilt $G_k = 2$, für Triphone gilt $G_k = 3$, usw. Für kontextabhängige Lauteinheiten gilt gemäß Gleichung (7) für den Kontextgrad

$$G_k = 1 + r + 1 \quad (8)$$

In der Erfindung wird zur Entscheidung, ob die Distanz einer Lauteinheit genauer (in einem höheren Detaillierungsgrad) berechnet werden soll, die Distanz der gleichen Lauteinheit im kleineren Kontextgrad (= abstrakteren Kontext) herangezogen. Das bedeutet, daß zur Entscheidung, ob die Distanz für die Lauteinheit

$$L_2(i-1, \dots, i-1, i+1, \dots, i+r)$$

explizit berechnet wird, die Distanz für einen abstrakteren Laut

$$L_{abs}(i-1_2, \dots, i-1, i+1, \dots, i+r_2)$$

herangezogen wird, wobei gilt $l_2 < l$, $r_2 < r$. Der Vorteil liegt darin, daß für alle $L_2(\dots)$, bei denen $l_2 < l$, $r_2 < r$ gilt, die Distanz der gleichen abstrakten Untereinheit L_{abs} zur Entscheidung verwendet werden kann. Demzufolge wird bei der Wahl der abstrakten Lauteinheit bis auf die Basis-Lauteinheiten zurückgegriffen, für die gilt $l=r=0$.

Sodann wird die optimale Distanz d_{opt} aus allen Distanzen der abstrakten Lauteinheiten berechnet (siehe Fig. 1, Schritt 1c):

$$d_{opt} = \min_j \{d_{abs,i}\} \quad \forall (i \in \{1, \dots, N_{U,abs}\}) \quad (9),$$

wobei $N_{U,abs}$ eine Anzahl der abstrakten Lauteinheiten bezeichnet und N_{abs} die Anzahl der Verteilungen, die die $N_{U,abs}$ abstrakten Lauteinheiten modelliert, bezeichnet.

Im Schritt 1e werden für Lauteinheiten mit höherem Kontextgrad die Distanzen explizit berechnet, wobei die Bedingung

$$d_{u,abs} < d_{opt} + thr \quad (10)$$

vorschreibt, daß innerhalb eines vorgebbaren Bereichs (gekennzeichnet durch den vorgebbaren Wert thr) um das nach Gleichung (9) errechnete Minimum d_{opt} abstrakte Lauteinheiten $d_{u,abs}$ zu einer detaillierteren Klassifizierung des Merkmalsvektors herangezogen werden (Schritt 1d).

Für den vorgebbaren Wert thr kann gelten: $thr = const$, d. h. es wird eine Konstante durch Vorberechnung bestimmt, oder der Wert thr wird implizit abhängig von dem augenblicklichen Merkmalsvektor x ausgewählt gemäß

$$thr = K_{thr} \times d_{opt} \quad (11),$$

wobei K_{thr} ein vorgebbarer Wert ist.

Für alle Lauteinheiten, deren Distanz der abstrakten Lauteinheit nach Gleichung (11) nicht innerhalb des durch den Wert thr vorgebbaren Bereichs liegt, d. h. deren Distanz zu groß ist, wird gesetzt:

$$d(L_2(i-1, \dots, i-1, i+1, \dots, i+r)) = d(L_{abs}) \quad (12).$$

Somit wird die Distanz der abstrakten Untereinheit gewählt, wodurch die explizite Berechnung der Distanz eingespart wird.

Die Anzahl der explizit zu berechnenden Verteilungen wird mit N_t bezeichnet, wobei $N_t \leq N_v$ gilt. Werden nun abstrakte Lauteinheiten zur Entscheidungsfindung nach Gleichung (10) ausgewählt, die gleichzeitig Teil des Spracherkennungssystems sind, wie bspw. die Basis-Lauteinheiten, sind die Verteilungen der abstrakten Lauteinheiten N_{abs} eine Teilmenge der modellierenden Verteilungen

$$N_{abs} \subseteq N_v \quad (13).$$

Somit gilt:

$$N_{ges} = N_{abs} + N_t \leq N_v \quad (14).$$

Auf diese Weise ist immer sichergestellt, daß nie mehr Verteilungen berechnet werden müssen, als das eigentliche System beinhaltet, da die zur Entscheidung herangezogenen Lauteinheiten bzw. Verteilungen Teil des Systems sind.

Gemäß obiger Beschreibung wird also innerhalb des durch den Wert thr vorgegebenen Bereichs wiederum ein Minimum der Abstände d_{opt} ermittelt (Schritt 1f) und, wenn die Güte der Klassifikation ausreicht (Schritt 1d), das Verfahren beendet, indem Lauteinheiten zusammengesetzt werden (Schritt 1g) und somit Sprache erkannt wird. Ist hingegen die Güte der Klassifikation noch nicht ausreichend, so wird zu Schritt 1e gesprungen und das Verfahren iteriert.

In Fig. 2 wird beispielhaft die Klassifikation im zweidimensionalen Merkmalsraum (x_1, x_2) veranschaulicht. In Fig. 2a sind ein Merkmalsvektor MV , und drei Lauteinheiten $LE1$, $LE2$ und $LE3$ mit demselben Kontextgrad dargestellt. Zu jeder Lauteinheit $LE1$, $LE2$ und $LE3$ wird jeweils eine Distanz d_1 , d_2 und d_3 berechnet. Das Minimum aus den Distanzen, hier der Abstand d_1 , wird gemäß Gleichung (9) ausgewählt.

In Fig. 2b wird um das ausgewählte Minimum ein vorgebbarer Bereich gemäß Gleichung (10) gekennzeichnet durch $d_1 + thr$, bestimmt. Im Beispiel liegen außer der Lauteinheit $LE1$ noch weitere Lauteinheiten $LE1-1$, $LE1-2$, $LE1-3$ und $LE1-4$ mit gegenüber Lauteinheit $LE1$ höherem Kontextgrad innerhalb dieses vorgebbaren Bereichs.

Für jede dieser Lauteinheiten LE1-1, LE1-2, LE1-3 und LE1-4 wird jeweils die Distanz berechnet. Aus den resultierenden Distanzen d1-1, d1-2, d1-3 und d1-4 wird das Minimum berechnet (wieder nach Gleichung (9)). Das Verfahren kann fortgesetzt iteriert werden, indem um das ausgewählte Minimum d1-4 wieder ein vorgegebbarer Bereich bestimmt wird und innerhalb dieses Bereichs Distanzen zu Lauteinheiten mit höherem Kontextgrad als die Lauteinheit LE1-4 berechnet werden.

Literatur:

[1] G. Ruske: Automatische Spracherkennung: Methoden der Klassifikation und Merkmalsextraktion. Oldenbourg, München 1988, ISBN 3-486-20877-2, S. 1—10.

[2] Enrico Bocchieri: Vector Quantization for the Efficient Computation of Continuous Density Likelihoods, Proc. IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP), S. II-692 bis II-695.

[3] L. R. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. IEEE, Feb. 1989, S. 257—286.

[4] L. R. Rabiner, B.-H. Juang, S. E. Levinson, M. M. Sondhi: Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities, AT&T Technical Journal, Vol.64, No.6, July-August 1985, S. 1211—1234.

Patentansprüche

1. Verfahren zur Spracherkennung durch einen Rechner,

- a) bei dem ein Sprachsignal mit einer vorgebbaren Taktrate digitalisiert wird,
- b) bei dem das digitalisierte Sprachsignal zu je einem Abtastzeitpunkt, der durch die Taktrate festgelegt ist, in einen Merkmalsvektor abgebildet wird,
- c) bei dem erste Abstände des Merkmalsvektors zu ersten Mustervektoren ermittelt werden, wobei die Mustervektoren jeweils eine kleinste unterscheidbare Lauteinheit mit einem vorgebbaren Kontextgrad, der angibt, wie weit der Kontext einer Lauteinheit reicht, darstellen,
- d) bei dem ein erstes Minimum der ersten Abstände ausgewählt wird,
- e) bei dem um das erste Minimum ein vorgebbbarer Bereich ausgewählt wird und in diesem Bereich zweite Abstände des Merkmalsvektors von zweiten Mustervektoren, die jeweils Lauteinheiten mit gegenüber den kleinsten unterscheidbaren Lauteinheiten höherem Kontextgrad darstellen, bestimmt werden,
- f) bei dem ein zweites Minimum der zweiten Abstände ausgewählt wird,
- g) bei dem der Merkmalsvektor durch die über das zweite Minimum erreichbare Lauteinheit klassifiziert wird oder, falls eine genauere Klassifikation erforderlich ist, das zweite Minimum als das erste Minimum angenommen und mit Schritt e) fortgefahren wird,
- h) bei dem aus zusammengesetzten klassifizierten Lauteinheiten Sprache erkannt wird.

2. Verfahren nach Anspruch 1, bei dem im Schritt e) für die zweiten Mustervektoren der nächsthöhere Kontextgrad gegenüber den ersten Mustervektoren ausgewählt wird.

3. Verfahren nach Anspruch 1 oder 2, bei dem die ersten Mustervektoren kontextunabhängige Einheiten mit dem Kontextgrad 1 darstellen.

4. Verfahren nach einem der Ansprüche 1 bis 3, bei dem als Mustervektoren Phoneme einer Sprache verwendet werden.

5. Verfahren nach einem der Ansprüche 1 bis 3, bei dem als Mustervektoren Triphone einer Sprache verwendet werden.

6. Verfahren nach einem der Ansprüche 1 bis 3, bei dem als Mustervektoren Quinphone einer Sprache verwendet werden.

Hierzu 2 Seite(n) Zeichnungen

FIG 2a

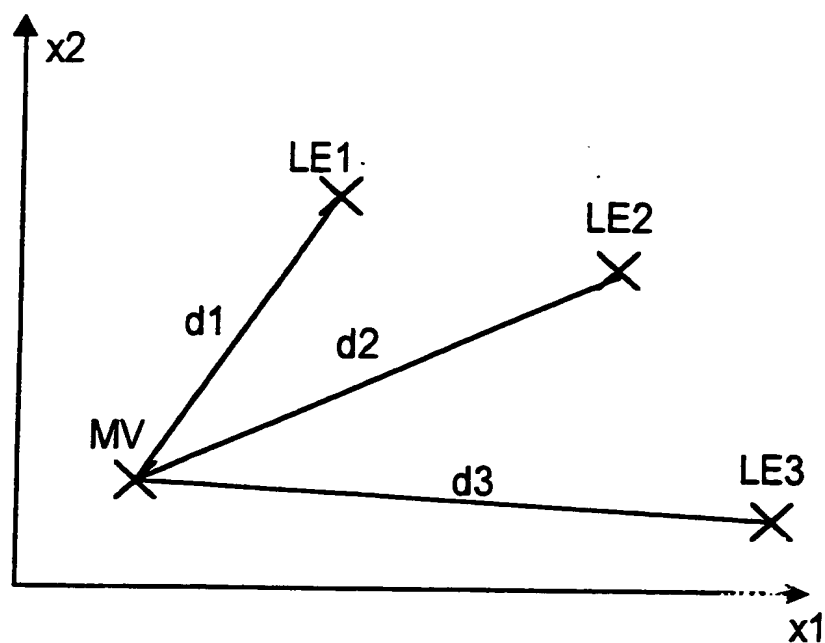


FIG 2b

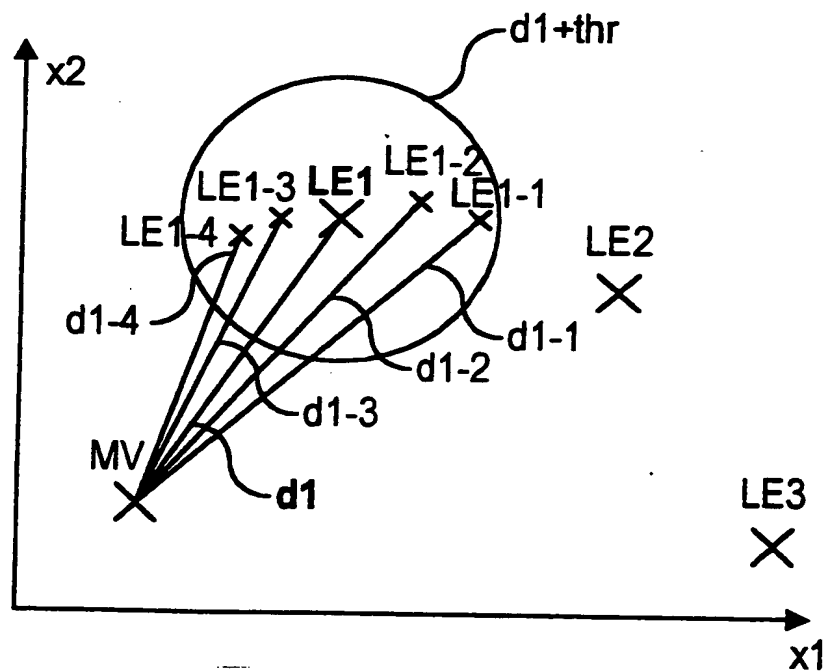


FIG 1

